CALCOLO 38, 97 – 112 (2001)

CALCOLO

Stopping criteria for iterative methods: applications to PDE's

M. Arioli[1], **E. Noulard**[2], **A. Russo**[1]

[1] Istituto di Analisi Numerica del C.N.R., Via Ferrata 1, 27100 Pavia, Italy
e-mail: arioli@dragon.ian.pv.cnr.it; russo@dragon.ian.pv.cnr.it
[2] Laboratoire d'Informatique et de Mathématiques Appliquées de l'ENSEEIHT, 2 rue Charles Camichel, 31071 Toulouse Cedex, France
e-mail: noulard@enseeiht.fr

Received: March 2000 / Accepted: October 2000

Abstract. We show that, when solving a linear system with an iterative method, it is necessary to measure the error in the space in which the residual lies. We present examples of linear systems which emanate from the finite element discretization of elliptic partial differential equations, and we show that, when we measure the residual in $H^{-1}(\Omega)$, we obtain a true evaluation of the error in the solution, whereas the measure of the same residual with an algebraic norm can give misleading information about the convergence. We also state a theorem of functional compatibility that proves the existence of perturbations such that the approximate solution of a PDE is the exact solution of the same PDE perturbed.

1 Introduction

Stationary physical phenomena are often driven by elliptic partial differential equations. The discretization of equations of this kind often leads to a real $N \times N$ linear system, $A \cdot \underline{x} = \underline{b}$, which is normally solved by Krylov-based methods such as Conjugate Gradient ([8]) when A is symmetric positive definite or GMRES ([12]) in the general case. At each iteration step we compute an approximation $\underline{x}^{(n)} \in \mathbb{R}^N$ of the solution of the linear system. It is necessary, at this point, to introduce a stopping criterion in order to test whether $\underline{x}^{(n)}$ is accurate enough for our purposes.

This work was supported by the "Istituto di Analisi Numerica – Consiglio Nazionale delle Ricerche" (Pavia, Italy) through the European programme HCM, contract no: ERBCHRXCT930420.

In previous reports ([9,2]) stopping criteria based on a backward error analysis of the algebraic problem were presented. In those reports, the algebraic residual $\underline{\rho}^{(n)} = A \cdot \underline{x}^{(n)} - \underline{b}$ was computed. By using a norm of the latter, it is possible to test whether the norm of the perturbations, for which $\underline{x}^{(n)}$ is the exact solution of a perturbed version of the original linear system, is sufficently small.

In the present paper, we focus on the fact that, when considering Galerkin-type discretizations of partial differential equations, the residual $\underline{\rho}^{(n)}$ defined above is the discrete counterpart of a linear functional, $R^{(n)}$, which belongs to the dual of the space that contains the exact solution. In particular, we want to show the difference between the algebraic and the functional convergence of a Krylov-based method when this method is applied to a linear system, which comes from the Galerkin discretization of an elliptic partial differential equation. We present the advantages of measuring the residual in the correct norm.

In Sect. 2, we define the abstract variational problem and its Galerkin discretization. In Sect. 3, we consider an elliptic partial differential equation in divergence form, and with a solution defined in the Sobolev space $H_0^1(\Omega)$. We introduce the measure of the residual on the corresponding dual space $(H_0^1(\Omega))' = H^{-1}(\Omega)$.

In Sect. 4, we describe a perturbation theory, in functional spaces, which generalizes that of Rigal and Gaches ([11]). In particular, we introduce a functional backward error analysis in such a way that an approximate finite dimensional solution may be considered as the exact solution of a perturbed version of the original continuous differential problem.

Finally, in Sects. 5 and 6, we describe practical aspects, test problems and numerical experiments.

2 The Galerkin framework

Suppose that we have a boundary value problem for a differential equation that can be set in the usual variational framework (a concrete example will be stated in Sect. 3). This means that we have a Hilbert space V, with a scalar product $(\cdot\,,\cdot)_V$, an induced norm $\|\cdot\|_V$, a bilinear form B on $V \times V$ and a linear form L on V with the following properties:

$$(\mathcal{H}) \begin{cases} B \text{ is continuous on } V \times V, \text{ i.e.,} \\ \exists M \text{ such that } \forall\, u, v \in V,\ B(u, v) \le M \|u\|_V \|v\|_V; \\ B \text{ is coercive, i.e.,} \\ \exists \gamma_B > 0, \forall\, u \in V,\ |B(u, u)| \ge \gamma_B \|u\|_V^2; \\ L \text{ is continuous on } V, \text{ i.e.,} \\ \forall\, u \in V,\ |L(u)| \le C \|u\|_V. \end{cases}$$

We denote by $\mathcal{BL}(V)$ the space of continuous bilinear forms $V \times V \to \mathbb{R}$, and by V' the topological dual space of V. Then V' is equipped with the classical dual norm:

$$\|F\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{|F(v)|}{\|v\|_V}.$$

The so-called variational formulation of the boundary value problem can then be written as follows:

$$(\mathcal{P}) \begin{cases} \text{find } u \in V \text{ such that} \\ B(u, v) = L(v) \; \forall \, v \in V. \end{cases}$$

It is well-known (Lax–Milgram Lemma) that, under assumption $(\mathcal{H})$, problem $(\mathcal{P})$ has a unique solution that depends continuously on the data (cf., e.g., [5]). The Galerkin discretization method consists in choosing a finite dimensional subspace V_h of V and then solving problem $(\mathcal{P})$ on V_h, i.e.,

$$(\mathcal{P})_h \begin{cases} \text{find } u_h \in V_h \text{ such that} \\ B(u_h, v_h) = L(v_h) \; \forall \, v_h \in V_h. \end{cases}$$

Problem $(\mathcal{P})_h$ still has a unique solution, because B and L, when restricted to V_h, satisfy properties $(\mathcal{H})$ (obviously we set $\|v_h\|_{V_h} = \|v_h\|_V \; \forall \, v_h \in V_h$). Let $\{\phi_i\}_{i=1,\dots,N}$ be a Lagrange basis for V_h. Then, $u_h = \sum_{j=1}^{N} u_j \phi_j$ and problem $(\mathcal{P})_h$ is equivalent to the linear system:

$$\sum_{j=1}^{N} B(\phi_j, \phi_i) u_j = L(\phi_i), \quad i = 1, \dots, N$$

which can be written:

$$\begin{aligned} & A \cdot \underline{x} = \underline{b}, \text{ where} \\ & A = (a_{ij})_{1 \le i \le N, 1 \le j \le N}, \text{ with } a_{ij} = B(\phi_j, \phi_i), \\ & \underline{x} = (x_j)_{1 \le j \le N}, \text{ with } x_j = u_j, \\ & \underline{b} = (b_i)_{1 \le i \le N}, \text{ with } b_i = L(\phi_i). \end{aligned}$$

If we solve the linear system with an iterative method, at step n we will have an approximate solution $\underline{x}^{(n)} = (x_j^{(n)}) \in \mathbb{R}^N$ with the corresponding $u_h^{(n)} \in V_h$ given by $u_h^{(n)} = \sum_j x_j^{(n)} \phi_j$.

3 A test problem

Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain. We consider the following elliptic boundary value problem (generally non-symmetric) which is defined in Ω:

$$\begin{cases} -\mathrm{div}(\nu \nabla u) + \boldsymbol{\beta} \cdot \nabla u + \alpha u = f \ \text{ in } \Omega \\ u = 0 \ \text{ on } \partial\Omega \end{cases} \tag{1}$$

where $\nu \in L^\infty(\Omega)$, $\boldsymbol{\beta} \in [C^1(\bar{\Omega})]^2$, $\alpha \in L^\infty(\Omega)$, $f \in L^2(\Omega)$. We assume the coerciveness hypotheses $0 < \nu_{\min} \leq \nu(x) \leq \nu_{\max} < +\infty$ and $-\frac{1}{2}\mathrm{div}\boldsymbol{\beta} + \alpha \geq 0$ pointwise. As usual, we denote by $H^1(\Omega)$ the Hilbert space of square-integrable functions, which are defined on Ω, with square-integrable first derivatives. Then $H^1(\Omega)$ is equipped with the scalar product and induced norm:

$$(u, v) = \int_\Omega uv + \int_\Omega \nabla u \nabla v$$

$$\|u\|_{1,\Omega} = (u, u)^{\frac{1}{2}} = \left(\int_\Omega u^2 + \int_\Omega |\nabla u|^2 \right)^{\frac{1}{2}}.$$

We then set

$$H_0^1(\Omega) = \left\{ v \in H^1(\Omega),\, v_{|\partial\Omega} = 0 \right\}.$$

It is well-known that the semi-norm on $H^1(\Omega)$ given by

$$|u|_{1,\Omega} = \left(\int_\Omega |\nabla u|^2 \right)^{\frac{1}{2}}$$

is indeed a norm on $H_0^1(\Omega)$ (Poincaré Lemma). In this way $H_0^1(\Omega)$ is a Hilbert space with scalar product

$$(u, v) = \int_\Omega \nabla u \nabla v$$

and induced norm $|u|_{1,\Omega}$. We denote by $H^{-1}(\Omega)$ the topological dual space of $H_0^1(\Omega)$. Problem (1) can then be written, in variational form, as follows:

$$(\mathcal{P}) \begin{cases} \text{find } u \in H_0^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nu \nabla u \cdot \nabla v + \int_\Omega (\boldsymbol{\beta} \cdot \nabla u) v + \int_\Omega \alpha u v = \int_\Omega f v, \quad \forall\, v \in H_0^1(\Omega), \end{cases}$$

where, referring to the previous section, we have $V = H_0^1(\Omega)$,

$$B(u, v) = \int_\Omega \nu \nabla u \cdot \nabla v + \int_\Omega (\boldsymbol{\beta} \cdot \nabla u) v + \int_\Omega \alpha u v,$$

and $L(v) = \int_\Omega f v$. Using the hypotheses stated above it is readily shown that B is continuous and coercive on $H_0^1(\Omega)$ and $L \in H^{-1}(\Omega)$. A finite element discretization of problem $(\mathcal{P})$ with the use of continuous piecewise linear elements can be described briefly as follows (for the sake of simplicity, we assume that all integrals are computed exactly). Let $\mathcal{T}_h$ be a family of triangulations of Ω, i.e., each $\mathcal{T}_h$ is a set of disjoint triangles $\{T\}$ which covers Ω in such a way that no vertex of any triangle lies in the interior of an edge of another triangle. We assume that $\mathcal{T}_h$ is compatible with the polygonal boundary of Ω. Let $h = \max_{T \in \mathcal{T}_h} \operatorname{diam}(T)$. Consider then the space

$$V_h = \left\{ v : \Omega \to \mathbb{R},\, v \in C^0(\overline{\Omega}),\, v_{|\partial\Omega} = 0,\, v_{|T} \text{ is linear } \forall\, T \in \mathcal{T}_h \right\}.$$

Thus $V_h \subset H_0^1(\Omega)$. Next we describe the usual finite element basis for V_h. Let $\left\{P_j\right\}_{j=1,\ldots,N}$ be the set of internal vertices of $\mathcal{T}_h$ (i.e., we exclude the vertices lying on $\partial\Omega$). Then for all j, $1 \le j \le N$, we define the function $\phi_j \in V_h$ by

$$\phi_j(P_i) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

and then we extend it linearly on each triangle T. It is easy to show that $\left\{\phi_j\right\}_{j=1,\ldots,N}$ is a basis for V_h; hence, $\dim V_h = N$. The finite element approximation $(\mathcal{P})_h$ of problem $(\mathcal{P})$ is then

$$(\mathcal{P})_h \begin{cases} \text{find } u_h \in V_h \text{ such that:} \\ \displaystyle\int_\Omega \nu \nabla u_h \cdot \nabla v_h + \int_\Omega (\boldsymbol{\beta} \cdot \nabla u_h) v_h + \int_\Omega \alpha u_h v_h = \int_\Omega f v_h, \\ \forall\, v_h \in V_h. \end{cases}$$

As shown in the previous section, problem $(\mathcal{P})_h$ is equivalent to a linear system $A \cdot \underline{x} = \underline{b}$, with

$$a_{ij} = B(\phi_j, \phi_i) = \int_\Omega \nu \nabla \phi_j \cdot \nabla \phi_i + \int_\Omega (\boldsymbol{\beta} \cdot \nabla \phi_j) \phi_i + \int_\Omega \alpha \phi_j \phi_i$$

and $b_i = \int_\Omega f \phi_i$. If we use an iterative method, at each step we will have a vector $\underline{x}^{(n)} \in \mathbb{R}^N$, which in turn identifies a function $u_h^{(n)} = \sum_{i=1}^N x_i^{(n)} \phi_i \in V_h$ and a residual $R_h^{(n)} \in V_h'$ (V_h' is the topological dual space of V_h) which is defined by

$$R_h^{(n)}(v_h) = \int_\Omega \left(\nu \nabla u_h^{(n)} \cdot \nabla v_h + (\boldsymbol{\beta} \cdot \nabla u_h^{(n)}) v_h + \alpha u_h^{(n)} v_h - f v_h \right), \quad \forall\, v_h \in V_h.$$

3.1 Measuring the residual

We want to measure the norm of $R_h^{(n)}$ in V_h'. This can be done, for instance, by solving the discretization of the Poisson problem

$$\begin{cases} -\Delta\rho = R & \text{in } \Omega \\ \rho = 0 & \text{on } \partial\Omega \end{cases} \tag{2}$$

in the same space V_h. The Poisson problem (2) induces an isometry $R \mapsto \rho$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$; we will see that its discretization on V_h induces an isometry between V_h and V_h'. Let $R_h \in V_h'$ (for simplicity we omit the superscript $^{(n)}$), and let ρ_h be the solution of the following variational problem:

$$\int_\Omega \nabla\rho_h \nabla v_h = R_h(v_h) \;\forall\, v_h \in V_h. \tag{3}$$

Then

$$\|R_h\|_{V_h'} = \sup_{v_h \in V_h \setminus \{0\}} \frac{|R_h(v_h)|}{|v_h|_{1,\Omega}} = \sup_{v_h \in V_h \setminus \{0\}} \frac{\left|\int_\Omega \nabla\rho_h \nabla v_h\right|}{|v_h|_{1,\Omega}}. \tag{4}$$

Consequently, for $v_h = \rho_h$ in (4),

$$\|R_h\|_{V_h'} \geq \frac{|\rho_h|_{1,\Omega}^2}{|\rho_h|_{1,\Omega}} = |\rho_h|_{1,\Omega},$$

and from the Cauchy–Schwarz inequality in (4) we obtain

$$\|R_h\|_{V_h'} \leq \sup_{v_h \in V_h \setminus \{0\}} \frac{|\rho_h|_{1,\Omega}|v_h|_{1,\Omega}}{|v_h|_{1,\Omega}} = |\rho_h|_{1,\Omega},$$

and then

$$\|R_h\|_{V_h'} = |\rho_h|_{1,\Omega}.$$

The variational problem (3) is, in turn, equivalent to the linear system in $\mathbb{R}^N$ $\Phi \cdot \underline{\rho} = \underline{R}$, where $\Phi_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i$, $\underline{\rho} = (\rho_i) \in \mathbb{R}^N$ are the components of ρ_h with respect to the basis $\{\phi_i\}$ and $\underline{R} = (R_i)$, $i = 1, \ldots, N$, with $R_i = R_h(\phi_i)$. If $v_h \in V_h$, $v_h = \sum_{i=1}^N v_i\phi_i$, then

$$|v_h|_{1,\Omega}^2 = \int_\Omega |\nabla v_h|^2 = \sum_{i,j=1}^N v_i v_j \int_\Omega \nabla\phi_i \cdot \nabla\phi_j = (\Phi \cdot \underline{v}, \underline{v}).$$

Hence,

$$|\rho_h|_{1,\Omega}^2 = (\Phi \cdot \underline{\rho}, \underline{\rho}) = (\Phi \cdot \Phi^{-1} \cdot \underline{R}, \Phi^{-1} \cdot \underline{R}) = (\underline{R}, \Phi^{-1} \cdot \underline{R}).$$

Thus, from the computational point of view, we have to solve a linear system and compute a scalar product.

Remark 1 Using the previous arguments we can also prove that

$$\|L_h\|^2_{V_h'} = (\underline{b}, \Phi^{-1} \cdot \underline{b}),$$

where $L_h = L|_{V_h}$. Moreover, because Φ is symmetric, the following inequalities are satisfied

$$\|\Phi\|_2^{-1/2}\|\underline{R}\|_2 \leq \|R_h\|_{V_h'} \leq \|\Phi^{-1}\|_2^{1/2}\|\underline{R}\|_2, \tag{5}$$

and

$$\|\Phi\|_2^{-1/2}\|\underline{b}\|_2 \leq \|L_h\|_{V_h'} \leq \|\Phi^{-1}\|_2^{1/2}\|\underline{b}\|_2, \tag{6}$$

where $\|.\|_2$ denotes the usual Euclidean norm of a vector and the spectral algebraic norm of a matrix.

4 Theoretical study

In this section, we state some results in perturbation theory within an abstract setting. Those who have some knowledge of matrix perturbation theory will recognize equivalent results for the compatibility of the solution of a linear system (see [11,9]).

4.1 A functional perturbation of problem $(\mathcal{P})$

Let $\tilde{u} \in V$ be an approximation of the solution $u \in V$ of the problem $(\mathcal{P})$ stated in Sect. 1. We want to prove the following theorem.

Theorem 1 (Compatibility) *We have the following equivalence.*

$$\left.\begin{array}{l}\exists \delta B \in \mathcal{BL}(V), \exists \delta L \in V' \text{ such that:} \\ (B+\delta B)(\tilde{u}, v) = (L+\delta L)(v), \\ \forall\, v \in V, \text{ and} \\ \|\delta B\|_{\mathcal{BL}(V)} \leq \alpha, \|\delta L\|_{V'} \leq \beta.\end{array}\right\} \Leftrightarrow \left\{\begin{array}{l}\|\rho_{\tilde{u}}\|_{V'} \leq \alpha\|\tilde{u}\|_V + \beta, \\ \text{where } \rho_{\tilde{u}} \in V' \text{ is defined by} \\ \langle \rho_{\tilde{u}}, v\rangle_{V',V} = B(\tilde{u}, v) - L(v) \\ \forall\, v \in V.\end{array}\right.$$

Proof The proof will be given under the assumption that V is only a Banach space, thereby showing that the theorem holds even in a more general situation. For this reason, in this proof (and only here), we use the notation of duality pairs.

$\Rightarrow$: This is obvious.

$\Leftarrow$: We will build two perturbations of B and L, respectively δB and δL, such that

$$B(\tilde{u}, v) + \delta B(\tilde{u}, v) = L(v) + \delta L(v) \forall\, v \in V.$$

We set

$$\forall\, u \in V, \langle \rho_u, v\rangle_{V',V} = B(u,v) - L(v), \forall\, v \in V;$$

then $\rho_u \in V'$. We denote by $J_u \in (V')' = V''$ the element of the bi-dual of V which is associated to u in the canonical injection

$$\begin{aligned} J : V &\longrightarrow V_I'' \subset V'' \\ u &\longmapsto J_u \end{aligned}$$

defined by $\langle J_u,\, f\rangle_{V'',V'} = \langle f, u\rangle_{V',V}\ \forall\, f \in V'$. It is well-known that J is a linear isometry (see, e.g., [3, III.4, p. 39] or [13, XIX.7]). We then have

$$\begin{aligned} \|J_{\tilde u}\|_{V''} = \|\tilde u\|_V = \sup_{\|f\|_{V'}\le 1} \langle J_{\tilde u},\, f\rangle_{V'',V'} &= \sup_{\|f\|_{V'}\le 1} \langle f, \tilde u\rangle_{V',V} \\ &= \langle f_{\tilde u}, \tilde u\rangle_{V',V} \end{aligned}$$

for a certain $f_{\tilde u} \in V'$. One must keep in mind of the fact that, here, we cannot associate a vector $v \in V$ to $f_{\tilde u}$ unless V is reflexive. In other words we cannot find a $v \in V$ such that $\|f_{\tilde u}\|_{V'} = \langle f_{\tilde u}, v\rangle_{V',V}$, because $\|f_{\tilde u}\|_{V'}$ is a sup and not a max. It is a max if (and only if) V is reflexive (see [3, p. 4]). Now, as has been done for the perturbation of a system of linear equations ([11]), we define:

$$\delta B(u,v) = -\frac{\alpha}{\alpha\|\tilde u\|_V + \beta}\, \langle J_u,\, f_{\tilde u}\rangle_{V'',V'}\, \langle \rho_{\tilde u}, v\rangle_{V',V} \tag{7}$$

and

$$\delta L(v) = \frac{\beta}{\alpha\|\tilde u\|_V + \beta}\, \langle \rho_{\tilde u}, v\rangle_{V',V}\,. \tag{8}$$

It is obvious that δB is continuous and bilinear from $V \times V$ to $\mathbb{R}$, and that $\delta L \in V'$; an easy computation shows that

$$\begin{aligned} &\delta L(v) - \delta B(\tilde u, v) \\ &\quad = \left(\frac{\beta}{\alpha\|\tilde u\|_V + \beta} + \frac{\alpha}{\alpha\|\tilde u\|_V + \beta}\, \langle J_{\tilde u},\, f_{\tilde u}\rangle_{V'',V'}\right) \langle \rho_{\tilde u}, v\rangle_{V',V} = \langle \rho_{\tilde u}, v\rangle_{V',V} \end{aligned}$$

as required. Moreover, if we suppose that $\|\rho_{\tilde u}\|_{V'} \le \alpha\|\tilde u\|_V + \beta$, then obviously from formulæ (7) and (8) we have:

$$\|\delta B\|_{\mathcal{BL}(V)} \le \alpha, \quad \|\delta L\|_{V'} \le \beta. \qquad \square$$

Remark 2 If V is a reflexive Banach space, we can give a more significant form to the perturbation term δB. In fact, in this case, we can identify J_u and u and obtain from (7) that

$$\begin{aligned}\delta B(u, v) &= -\frac{\alpha}{\alpha\|\tilde{u}\| + \beta} \langle J_u, f_{\tilde{u}} \rangle_{V'',V'} \langle \rho_{\tilde{u}}, v \rangle_{V',V} \\ &= -\frac{\alpha}{\alpha\|\tilde{u}\| + \beta} \langle f_{\tilde{u}}, u \rangle_{V',V} \langle \rho_{\tilde{u}}, v \rangle_{V',V} \\ &= -\frac{\alpha}{\alpha\|\tilde{u}\| + \beta} \langle f_{\tilde{u}} \otimes \rho_{\tilde{u}}, (u, v) \rangle ,\end{aligned}$$

in analogy with the finite dimensional case (see, e.g., [9]).

4.2 Conditioning of $(\mathcal{P})$

In the theory of linear systems, we can define several condition numbers which are associated with the problem $A \cdot \underline{x} = \underline{b}$. These condition numbers measure "the numerical difficulty" we have in solving the linear system. The best known condition number is the condition number of the matrix A itself,

$$\kappa(A) = \|A\|_2 \cdot \|A^{-1}\|_2.$$

The reciprocal of $\kappa(A)$ measures the distance, in spectral norm, of A from the class of singular matrices (see [6]). Furthermore, the condition number of a problem can be defined as the least upper bound of the ratio of the norm of perturbation in the solution to the norm of perturbation in the input data, in the limit as the perturbation in the input data goes to zero (see, e.g., [1, 9] or [4]). Here we want to define a condition number for the problem $(\mathcal{P})$ as defined in Sect. 1.

Definition 1 (Condition number) *Let $\delta B \in \mathcal{BL}(V)$ and $\delta L \in V'$ be two perturbations of B and L such that*

$$\|\delta B\|_{\mathcal{BL}(V)} \leq \varepsilon\alpha, \quad \|\delta L\|_{V'} \leq \varepsilon\beta.$$

The relative condition number, $C(\mathcal{P})$, for the variational problem $(\mathcal{P})$ is the smallest constant C which satisfies the inequality

$$\|u - \tilde{u}\|_V \leq \varepsilon C \|u\|_V.$$

We have the following theorem.

Theorem 2 (Condition number) *The condition number $C(\mathcal{P})$ of definition 1 satisfies the bound*

$$C(\mathcal{P}) \leq \frac{(\alpha\|u\|_V + \beta)}{\gamma_B \|u\|_V}.$$

Proof

$$
\begin{aligned}
(B+\delta B)(\tilde{u},v) &= (L+\delta L)(v)\ \forall\, v\in V\\
&\Leftrightarrow B(u-\tilde{u},v) = -\delta L(v)+\delta B(\tilde{u},v)\ \forall\, v\in V.\\
v=u-\tilde{u} &\Rightarrow B(u-\tilde{u},u-\tilde{u}) = -\delta L(u-\tilde{u})+\delta B(\tilde{u},u-\tilde{u}).\\
B \text{ is coercive} &\Rightarrow |-\delta L(u-\tilde{u})+\delta B(\tilde{u},u-\tilde{u})| \geq \gamma_B\|u-\tilde{u}\|_V^2\\
&\Rightarrow \gamma_B\|u-\tilde{u}\|_V^2 \leq \|\delta L\|_{V'}\|u-\tilde{u}\|_V\\
&\qquad\qquad + \|\delta B\|_{\mathcal{BL}(V)}\|u-\tilde{u}\|_V\|\tilde{u}\|_V.\\
u-\tilde{u}\neq 0 &\Rightarrow \gamma_B\|u-\tilde{u}\|_V \leq \|\delta L\|_{V'}\\
&\qquad\qquad + \|\delta B\|_{\mathcal{BL}(V)}\left(\|u\|_V+\|u-\tilde{u}\|_V\right)\\
&\Rightarrow (\gamma_B-\varepsilon\alpha)\|u-\tilde{u}\|_V \leq \varepsilon\left(\beta+\alpha\|u\|_V\right).\\
\text{If } \varepsilon\alpha<\gamma_B &\Rightarrow \|u-\tilde{u}\|_V \leq \varepsilon\left(1-\varepsilon\tfrac{\alpha}{\gamma_B}\right)^{-1}\tfrac{1}{\gamma_B}\left(\alpha\|u\|_V+\beta\right). \qquad \square
\end{aligned}
$$

4.3 Practical consequences

It is easy to see that all the inequalities shown before hold even if we work in V_h and V_h' instead of in V and V', and the constants involved do not depend on V_h. This means that we have defined a "functional" condition number, which is independent of the discretization. This may seem strange at first glance, but it depends on the choice of the norms which are used to measure the residual. For instance, in the discretization of the Laplace operator, the "classical" condition number of the stiffness matrix is proportional to $1/h^2$, while our "functional" condition number is always constant. It then seems reasonable to consider

$$\frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}}$$

as the relative functional backward error on u_h. Assume that

$$\frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}} \leq \varepsilon.$$

According to the above definition of the condition number, from the previous theorem (with $\alpha=0$ and $\beta=\|L_h\|_{V_h'}$) we have

$$\frac{\|u_h-u_h^{(n)}\|_{V_h}}{\|u_h\|_{V_h}} \leq \varepsilon\cdot\frac{\|L_h\|_{V_h'}}{\gamma_B\|u_h\|_{V_h}}.$$

Since

$$B(u_h,\cdot)=L_h \text{ in } V_h',$$

we have

$$\|L_h\|_{V_h'} \leq M\|u_h\|_{V_h}$$

and then the following relative forward error bound holds:

$$\frac{\|u_h - u_h^{(n)}\|_{V_h}}{\|u_h\|_{V_h}} \leq \varepsilon M {\gamma_B}^{-1}.$$

4.4 Stopping criteria and approximation error

We next propose a reasonable threshold for the relative functional backward error defined above. Let $u \in H_0^1(\Omega)$ be the exact solution of problem $(\mathcal{P})$, $u_h \in V_h$ its Galerkin approximation and $u_h^{(n)} \in V_h$ the algebraic approximation of u_h, computed with either the Conjugate Gradient or the GMRES method. Then let $R_h^{(n)}$ be the residual functional associated with $u_h^{(n)}$.

Moreover, we know an *a priori* bound for the functional approximation error. For instance, when using linear finite elements and under some mild assumptions on the triangulation $\mathcal{T}_h$, we have (see, for example, [5] and [10, pp. 110–111, Corollaire 5.1–3.])

$$|u_h - u|_{1,\Omega} \leq Ch|u|_{2,\Omega},$$

where the constant C depends only on the domain and on the coefficients of the equation.

In order to bound the global error $|u_h^{(n)} - u|_{1,\Omega}$, we can add and subtract u_h and bound separately the functional algebraic error and the functional approximation error:

$$\begin{aligned} |u_h^{(n)} - u|_{1,\Omega} &\leq |u_h - u|_{1,\Omega} + |u_h^{(n)} - u_h|_{1,\Omega} \\ &\leq Ch|u|_{2,\Omega} + \|u_h\|_{V_h} M {\gamma_B}^{-1}\varepsilon. \end{aligned}$$

Thus, it seems reasonable to stop the iteration when the upper bound of the functional algebraic error becomes smaller than the upper bound of the functional approximation error. Asymptotically, this can be achieved by the following stopping criteria:

$$\frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}} \approx h^2. \tag{9}$$

Finally, in several practical problems the linear form $\|L\|_{V'}$ is only experimentally determined. Therefore, it is perturbed by experimental errors of which we know an upper bound for the functional norm.

In these situations, it can be sensible to substitute in (9) the former h^2 with the latter upper bound for the experimental error functional norm.

5 Computational aspects

Measuring the norm of the residual in the way described above is interesting, but it should not force us to spend too much time in computing it. In this section, we briefly consider the cost of this computation. Note that the implementation cost of our criteria depends only on the solution of the isometry operator. For symmetric elliptic problems the symmetric bilinear form $B(\cdot,\cdot)$, ($B(u,v) = B(v,u)$), induces a norm equivalent to $\|\cdot\|_V$ and a isometry between V and V'. In these cases the corresponding linear system is solved using Conjugate Gradient since the matrix A is symmetric and positive definite. Because Conjugate Gradient minimizes at each step the dual norm of the residual, $(\underline{R}^{(n)}, A^{-1}\underline{R}^{(n)})^{1/2}$, on a Krylov subspace, it is quite appropriate to use the results of [7] to evaluate this norm directly. Frequently, a Krylov approximation method does not provide the residual but the preconditioned residual $P^{-1}\cdot(A\cdot\underline{x}^{(n)} - \underline{b})$, where P is an algebraic preconditioner. One should be aware of the fact that applying the isometry (Δ_h^{-1}, or A^{-1} if A is symmetric and positive definite) to this algebraic preconditioned residual may be senseless if it does not correspond to a functional residual in H^{-1}. If this occurs, we must compute $\underline{\rho} = A\cdot\underline{x}^{(n)} - \underline{b}$, which costs an extra matrix/vector product.

In the following, we experiment only with unsymmetric operators.

When we dispose of the residual, we have to solve a Poisson problem on the mesh which we built for our approximations.

- If our mesh is regular, we can use the FFT to solve the Poisson problem.
- If our mesh is not regular, the problem is more complex and deserves further investigation. There are several possibilities: interpolate on a regular grid and then use FFT; solve the Poisson problem at selected iterations; etc.
- We can use a few iterations of the Conjugate Gradient algorithm to reach a reasonable approximation of $(\underline{R}, \Phi^{-1}\cdot\underline{R})$ as is shown in [7].

Another and more practical way of bounding the functional residual is the following. Using the inequalities (5) and (6) we have

$$\kappa(\Phi)^{-1/2}\frac{\|\underline{R}^{(n)}\|_2}{\|\underline{b}\|_2} \le \frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}} \le \kappa(\Phi)^{1/2}\frac{\|\underline{R}^{(n)}\|_2}{\|\underline{b}\|_2}.$$

Because the algebraic condition number of the matrix Φ (discretizing the Poisson operator) is of order h^{-2}, we have

$$\mathcal{O}(h)\frac{\|\underline{R}^{(n)}\|_2}{\|\underline{b}\|_2} \le \frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}} \le \mathcal{O}(h^{-1})\frac{\|\underline{R}^{(n)}\|_2}{\|\underline{b}\|_2}.$$

Therefore, we can use the following as a rough stopping criterion:

– select $\epsilon \approx h^2$;
– **IF** $\frac{\|\underline{R}^{(n)}\|_2}{\|\underline{b}\|_2} \leq \epsilon$ **THEN**
IF $\frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}} > \epsilon$ **THEN** $\epsilon = h\epsilon$ **ELSE** *STOP*.

This stopping criterion can be implemented using the functional residual only once, with a reasonable extra cost if we use the modified version of the Conjugate Gradient algorithm proposed in [7]. We must point out that this criterion can be misleading when

$$\frac{\|R_h^{(n)}\|_{V_h'}}{\|L_h\|_{V_h'}} = \mathcal{O}(h^{-1})\frac{\|\underline{R}^{(n)}\|_2}{\|\underline{b}\|_2}.$$

For 1-D problems, this can be the case when

$$f(x) = \sin(N\pi x)$$

and $R_h^{(n)}$ is a smooth function. Nevertheless, this is not a common situation: it is more common to have the opposite where f is a regular and $R_h^{(n)}$ has the erratic behavior typical of an irregular function. In this last case, the proposed stopping criterion can force us to do more iterations than necessary.

6 Numerical experiments

In this section, we present some numerical experiments to illustrate the main ideas of the paper. In all cases, the domain Ω is the square $[0, 1] \times [0, 1]$ which is discretized with the mesh shown in Fig. 1 with $h \simeq 1/20$. The right-hand side is always $f(x, y) = 1$.

For each test case, we follow the convergence history of the following quantities:

– the Euclidean norm of the (algebraic) residual scaled with the right-hand side, i.e., $\frac{\|\underline{\rho}^{(n)}\|_2}{\|\underline{b}\|_2}$ *(solid)*
– the functional norm of the (functional) residual scaled with the right-hand side, i.e., $\frac{\|R_h^{(n)}\|_{V_h'}}{\|L\|_{V_h'}}$ *(dotted)*
– the norm of the forward relative error in V_h, i.e., $\frac{\|u_h^{(n)} - u_h\|_{V_h}}{\|u_h\|_{V_h}}$ *(dashed)*.

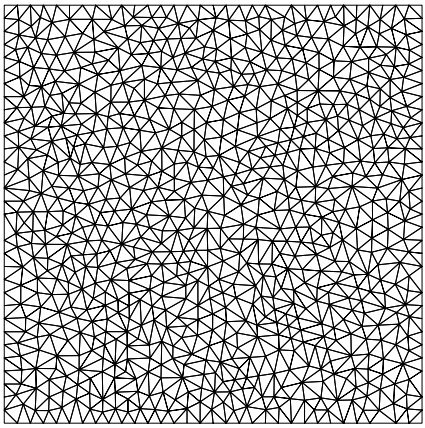

Fig. 1. Mesh; 1601 elements, 863 nodes (740 internal), $h \simeq 1/20$

In our experiments, we considered only the convection-diffusion equation. Naturally, as far as the ratio diffusion/convection is of the same order (or larger) than h, the problem is diffusion-dominated and it behaves like a pure diffusive case. When the ratio diffusion/convection becomes smaller, then the problem becomes convection-dominated and the Galerkin method described above produces a solution with spurious oscillations. In this case, a *stabilized* method is needed, and the norms involved change. In the experiments below, the values of ν have always been chosen in such a way that the problem is diffusion-dominated.

For this case we have used the GMRES method without restarting (and with no preconditioner). We have taken a fixed convection $\boldsymbol{\beta} = (1, 3)$ with $\nu = 0.1$ (Fig. 2) and $\nu = 0.01$ (Fig. 3). Also in this case we see that the algebraic norm of the residual is an overestimate of the exact error.

7 Conclusion

We have shown that, in the iterative solution of linear systems which arise from the Galerkin discretization of elliptic partial differential equations, the stopping criteria should rely on the H^{-1} norm residual measure.

This norm gives precise information about the true error, i.e., the error between the computed solution and the exact solution of the partial differential equation.

Moreover, we show that the threshold in our stopping criterion can be related to the value of the discretization error so that the algebraic part of the error can be safely compared with the discretization error, and we can stop when these two errors are of the same order. Some simple numerical experiments show that, if the dual norms can be computed, then we stop

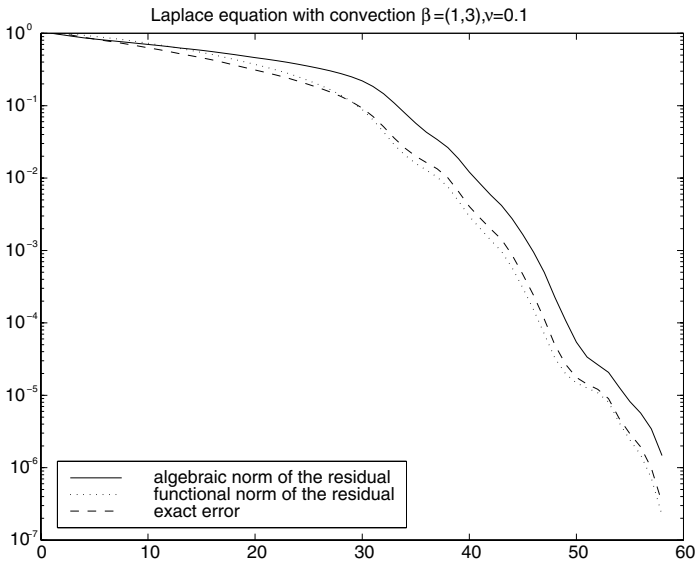

Fig. 2. Convection-diffusion equation with $\boldsymbol{\beta} = (1, 3)$, $\nu = 0.1$

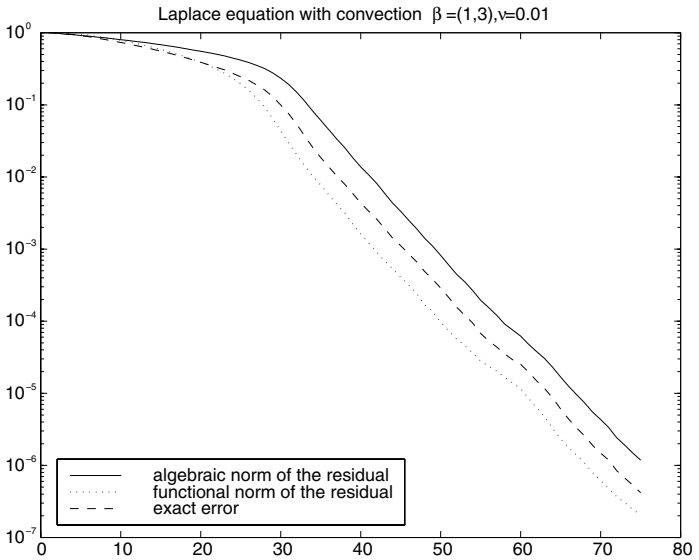

Fig. 3. Convection-diffusion equation with $\boldsymbol{\beta} = (1, 3)$, $\nu = 0.01$

when the Euclidean norm of the algebraic residual is still very high, but with a very satisfactory solution. Naturally, the drawback is that dual norms are not always cheap to compute. Nevertheless, we are confident that in some cases the computational cost can be reduced to an acceptable level.

References

[1] Arioli, M., Demmel, J.W., Duff I.S.: Solving sparse linear systems with sparse backward error. SIAM J. Matrix Anal. Appl. **10**, 165–190 (1989)
[2] Arioli, M., Duff, I.S., Ruiz, D.: Stopping criteria for iterative solvers. SIAM J. Matrix Anal. Appl. **13**, 138–144 (1992)
[3] Brezis, H.: Analyse fonctionnelle. Théorie et applications. Paris: Masson 1983
[4] Chaitin-Chatelin, F., Frayssé, V.: Lectures on finite precision computations. Philadelphia: SIAM 1995
[5] Ciarlet, P.G.: The finite element method for elliptic problems. Amsterdam: North-Holland 1978
[6] Demmel, J.W.: On the condition numbers and the distance to the nearest ill-posed problem. Numer. Math. **51**, 251–289 (1987)
[7] Golub, G.H., Meurant, G.: Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. BIT **37**, 687–705 (1997)
[8] Greenbaum, A.: Iterative methods for solving linear systems. Philadelphia: SIAM 1997
[9] Noulard, E., Arioli, M.: Vector stopping criteria for iterative methods: Theoretical tools. IAN-CNR Rep. No. **956**. Pavia: Istituto di Analisi Numerica - C.N.R. Pavia: 1995; Tech. Rep. LIMA-ENSEEIHT RT/APO/95/4. Toulouse: Laboratoire Informatique et Mathématiques Appliquées
[10] Raviart, P.-A., Thomas, J.-M.: Introduction à l'analyse numérique des équations aux dérivées partielles. Paris: Masson 1983
[11] Rigal, J.-L., Gaches, J.: On the compatibility of a given solution with the data of a linear system. J. Assoc. Comput. Mach. **14**, 543–548 (1967)
[12] Saad, Y., Schultz, M.H.: GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Statist. Comput. **7**, 856–869 (1986)
[13] Schwartz, L.: Topologie générale et analyse fonctionnelle. 2ième ed., Paris: Hermann 1986